What is the impact of incorporating explicit prosodic feature representations (e.g., pitch, energy, rhythm) on

Assignee Research

June 10, 2026

Abstract

Speech inherently contains rich acoustic information that extends far beyond the textual language. In real-world spoken language understanding, effective interpretation often requires integrating semantic meaning (e.g., content), paralinguistic features (e.g., emotions, speed, pitch) and phonological characteristics (e.g., prosody, intonation, rhythm), which are embedded in speech. While recent multimodal Speech Large Language Models (SpeechLLMs) have demonstrated remarkable capabilities in processing audio information, their ability to perform fine-grained perception and complex reasoning in

1 Introduction

This paper examines: MMSU: A Massive Multi-task Spoken Language Understanding and Reasoning Benchmark. Research question: What is the impact of incorporating explicit prosodic feature representations (e.g., pitch, energy, rhythm) on the downstream performance of multimodal LLMs, as measured by improvements in accuracy or F1 scores on the MMSU prosody-related tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

12 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
MMSU encompasses a wider range of acoustic features spanning 47 distinct tasks.	×	0.09
MMSU is the first benchmark to systematically incorporate linguistically grounded phenomena into spoken language underst	✓	0.21
MMSU requires models to integrate paralinguistic, phonetic, and semantic information for tasks such as sarcasm detection	×	0.09
MMSU includes 47 distinct tasks covering various linguistic phenomena and acoustic features.	×	0.11
MMSU is evaluated on 22 models, including 12 Speech-LLMs and 10 Omni Large Language Models (OmniLLMs) with audio process	×	0.08
The evaluation strategy for MMSU involves an audio clip and a text prompt, with the model choosing one of four options (	×	0.04
Answer options in MMSU are randomly ordered and balanced across the dataset to avoid positional bias.	×	0.02
All models in MMSU are evaluated with the same optimized instruction-following prompts to ensure fairness and minimize p	×	0.05
The question 'What is the intonation of the entire sentence in the audio?' has options: (A) Failing Intonation, (B) Risi	×	0.03
The correct answer to the question 'What is the intonation of the entire sentence in the audio?' is (A) Failing Intonati	×	0.05
Qwen2.5-Omni-7B incorrectly answered the question 'What is the intonation of the entire sentence in the audio?' with (D)	×	0.03

References

- http://arxiv.org/abs/2606.05868v1
- http://arxiv.org/abs/2306.13394v5
- http://arxiv.org/abs/2506.04779v3